NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Text Mining e Twitter:
O poder das redes sociais num mercado competitivo

Hugo Miguel Ferrão Casal da Veiga

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

TEXT MINING E TWITTER:
O PODER DAS REDES SOCIAIS NUM MERCADO COMPETITIVO

por

Hugo Miguel Ferrão Casal da Veiga

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do Conhecimento e Business Intelligence

Orientador: Professor Doutor Roberto Henriques

Co-orientador: Mestre Ivo Bernardo

Novembro 2015

AGRADECIMENTOS

Gostaria de agradecer ao meu orientador e co-orientador por me guiarem no desenvolvimento deste projecto. À NOVA IMS, enquanto entidade formadora, que me proporcionou uma educação de elevada qualidade e foi como uma segunda casa para mim. Gostaria também de agradecer à família e amigos, em especial ao Frederico e à Catarina, pelo apoio imprescindível prestado. Por fim, mas não menos importante, agradecer à minha avó pela minha formação enquanto pessoa. Sem ela o caminho percorrido até aqui, e culminado neste trabalho, não teria sido possível. A todos, obrigado.

RESUMO

Actualmente, com a massificação da utilização das redes sociais, as empresas passam a sua mensagem nos seus canais de comunicação, mas os consumidores dão a sua opinião sobre ela. Argumentam, opinam, criticam (Nardi, Schiano, Gumbrecht, & Swartz, 2004). Positiva ou negativamente. Neste con*exto* o *Text Mining* surge como uma abordagem interessante para a resposta à necessidade de obter conhecimento a partir dos dados existentes. Neste trabalho utilizámos um algoritmo de *Clustering* hierárquico com o objectivo de descobrir temas distintos num conjunto de *tweets* obtidos ao longo de um determinado período de tempo para as empresas Burger King e McDonald's. Com o intuito de compreender o *sentimento* associado a estes temas foi feita uma análise de *sentimentos* a cada tema encontrado, utilizando um algoritmo *Bag-of-Words*. Concluiu-se que o algoritmo de *Clustering* foi capaz de encontrar temas através do *tweets* obtidos, essencialmente ligados a produtos e serviços comercializados pelas empresas. O algoritmo de *Sentiment Analysis* atribuiu um *sentimento* a esses temas, permitindo compreender de entre os produtos/serviços identificados quais os que obtiveram uma polaridade positiva ou negativa, e deste modo sinalizar potencias situações problemáticas na estratégia das empresas, e situações positivas passíveis de identificação de decisões operacionais bem-sucedidas.

PALAVRAS-CHAVE

Big Data; *Clustering; Sentiment Analysis* ; *Text Mining*; Twitter

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

INTRODUÇÃO

A criação de informação a nível mundial cresce a um ritmo elevado. Em 2012 foram gerados 2,72 *zettabytes* de informação (1 *zettabyte* = Mil milhões de *terabytes*) a nível mundial, prevendo-se que sejam gerados 8 *zettabytes* em 2015 (Sagiroglu & Sinanc, 2013). Vivemos na era do *Big Data*. O *Big data* é caracterizado por 3 "V's". **Variedade** (grandes quantidades de dados, não estruturados, estruturados ou semi-estruturados) que necessitam de uma grande capacidade de processamento e armazenamento por parte de quem a gere, (dados não estruturados são representados na forma de *text*o, vídeo, entre outros), **Volume**, representando o tamanho avultado dos dados gerados, e a **Velocidade** elevada a que estes dados são produzidos (Sagiroglu & Sinanc, 2013). Este novo paradigma leva-nos a colocar algumas questões. Terão estes dados relevância a nível organizacional? Deverão as empresas dar relevância a estes dados por forma a aumentar a satisfação e retenção dos seus clientes? A polaridade da informação partilhada pelos clientes/utilizadores (positiva ou negativa) deverá ser tida em conta?

Tumasjan, Sprenger, Sandner, & Welpe (2010) utilizando o Twitter, concluíram que esta plataforma foi utilizada como um meio de deliberação política, aquando das eleições Federais para o Parlamento nacional Alemão em 2009, tendo sido encontrada correspondência entre as posições políticas dos partidos e o *sentimento* dos posts dos utilizadores. He, Zha, & Li (2013), utilizando dados do Twitter e Facebook nas análises, realizaram um estudo de análise competitiva sobre empresas líderes no mercado de pizarias norte-americanas. Utilizaram dados do Twitter e Facebook nas análises, recorrendo a técnicas de *Text Mining*. Neste estudo, os autores concluíram que as empresas que analisaram conseguiram não só promover-se nos seus canais sociais, mas também criar uma relação de proximidade com os seus seguidores, sugerindo que as redes sociais têm um papel importante na gestão da relação com os clientes. Estes estudos indicam que a utilização das redes sociais pode de facto ter uma relação interessante e útil com o mundo "real".

Tendo em conta o explicitado, decidimos estudar uma rede social (no caso o Twitter) e a ligação que a mesma poderia ter com o mundo empresarial. Escolhemos duas empresas concorrentes, intervenientes no sector da restauração, reconhecidas mundialmente, com uma presença física global e presença na rede social escolhida. As empresas escolhidas foram o Burger King e o McDonald's. A página de Twitter do Burger King tem aproximadamente 1 milhão e 200 mil seguidores e a do McDonald's tem aproximadamente 3 milhões seguidores. Segundo o relatório publicado pela American Satisfaction Index (que produz um score de satisfação dos clientes para as empresas norte-americanas que varia entre 0%-100%) com informação relativa ao primeiro semestre de 2015 sobre o sector da restauração de serviço limitado (cadeias de fast food) norte-americano (American Customer Satisfaction Index, 2015), tanto o Burger King como o McDonald's apresentam quedas no indicador de satisfação dos seus clientes face ao ano anterior na ordem dos 4 pontos percentuais para o Burger King (de 76% para 72%) e de 4 pontos percentuais também para o McDonald's (de 71% para 67%) estando o Burger King colocado em 15ª lugar, o McDonald's colocado na última posição em 18 empresas analisadas, e apesar de ser a maior cadeia de fast food do mundo, apresenta quedas das vendas nos Estados Unidos da América pelo sexto trimestre consecutivo. O relatório aponta a retoma da economia americana, ainda que lenta, como um factor importante na mudança de escolhas alimentares dos consumidores. Se por um lado com a retracção da economia o preço acaba por ser mais importante que a qualidade, no cenário oposto a preferência por artigos

considerados mais frescos e saudáveis começa a ganhar peso, mesmo que isso implique que os consumidores gastem um pouco mais. É importante referir que a média geral do sector é de 77% em 2015, tendo também sofrido uma queda face ao ano de 2014 (menos 3 pontos percentuais). Empresas mais recentes como a Chipotle Mexican Grill (especializada em comida mexicana) aparece em 2015 com uma pontuação de 83% directamente para o 2º lugar do ranking. O relatório justifica este facto com a disponibilização aos consumidores de produtos orgânicos, produtos não modificados geneticamente e opções mais saudáveis dizendo ainda que o mercado está cada vez mais competitivo e empresas jovens começam a ameaçar os grandes impérios de marcas reconhecidas e estabelecidas no último meio século. Tendo em conta este cenário do tecido empresarial da restauração norte-americana e as potencialidades do *Text Mining* aplicado aos dados das redes sociais, seguindo o Burger King e o McDonald's, com o presente projecto propomo-nos a:

- Apresentar os algoritmos de *Text Mining* que iremos aplicar sobre os dados extraídos do Twitter;
- Encontrar padrões nos dados que permitam compreender o valor da informação proveniente do Twitter;
- Analisar os modelos implementados com vista à identificação de situações negativas e positivas que possam pôr em causa o sucesso de determinadas acções/decisões das empresas, e elaborar recomendações com base nos resultados;

Pretendemos com este estudo realçar a possibilidade de criação de valor para as empresas através da obtenção de conhecimento sobre os seus clientes por via das redes sociais. O objectivo será conseguir provar que, através do Twitter, é possível compreender os temas mais abordados pelos utilizadores em relação às empresas em estudo, num determinado período de tempo, e ao mesmo tempo perceber qual a polaridade associada a esses temas. Pretendemos desta forma conseguir fornecer orientações às empresas para que estas possam posteriormente ter uma melhor compreensão sobre os potenciais motivos que poderão ter levado a desalinhamentos entre resultados previstos e resultados obtidos para os seus produtos/serviços com base nas opiniões que as pessoas expressam no Twitter sobre os mesmos. No capítulo "Revisão da Literatura" apresentamos uma revisão de literatura existente sobre as temáticas do Twitter e *Text Mining*. No capítulo "Metodologia" apresentamos a metodologia desenvolvida para a obtenção e tratamento dos dados, bem como a base teórica que suporta os algoritmos utilizados na modelação. No capítulo "Resultados e Discussão" apresentamos os resultados obtidos para os modelos desenvolvidos e elaboramos algumas recomendações direccionadas às empresas.

REVISÃO DA LITERATURA

1.1. REDES SOCIAIS E TWITTER

Uma rede social pode ser definida (de forma simplista e dentro do con*text*o da Web 2.0) como uma plataforma online onde os utilizadores partilham informação e criam conexões com outras pessoas e entidades, criando uma rede, ainda que interajam apenas como uma parte dessa mesma rede (Huberman, Romero, & Wu, 2008). A Setembro de 2015 a rede social Facebook atingiu 1,55 mil milhões de utilizadores mensais activos (Facebook, 2015) e a rede social Twitter, em média, 307 milhões (Statista, 2015). Qualquer pessoa com acesso à internet tornou-se num criador e simultaneamente num consumidor de conteúdos. Akshay, Xiaodan, Tim, & Java, (2007), concluíram que os principais tipos de "intenções" de pessoas que utilizam o Twitter são conversas de circunstância, partilha de informação e reporte de notícias. Quer isto dizer que os utilizadores tornam-se veículos de informação, não só recebendo mas também partilhando. O facto de existir esta partilha torna as redes sociais num local de potencial geração de reconhecimento (bom ou mau) para as empresas, sendo que quando estas publicam conteúdos que os utilizadores acabam por considerar não positivo ou duvidoso, os impactos são negativos, tal como aconteceu no passado com empresas como a Vodafone, o Bing ou o McDonald's (Bhasin, 2012). Da mesma forma temos uma perspectiva positiva no exemplo da empresa Dell, que declarou no passado ter gerado 1 milhão de dólares de receita adicional proveniente de alertas de vendas colocados no Twitter (Kaplan & Haenlein, 2010), o que indica que uma boa gestão deste canal de comunicação pode trazer de facto valor às empresas.

Os dados para o estudo serão extraídos da rede social Twitter (www.twitter.com) cujo conceito consiste na partilha de mensagens de até 140 caracteres. Existe ainda uma funcionalidade que consiste na realização de posts (chamados de *tweets*) com a utilização do caracter "#" (conhecido como "hashtag") no início de cada post de modo a centralizar a informação por assunto.

Consideremos o seguinte exemplo de um *tweet*:

"Finalmente de volta ao meu país! @Joaorodrigues77 temos de combinar uma ida aos pastéis de Belém! #Portugal #devolta #pasteisdebelem" Publicado às 10:00 do dia 20/12/2012

Este *tweet* dá-nos informação sobre o utilizador, nomeadamente a hora e data em que o post foi feito, e pelo conteúdo do mesmo, que se encontra em Portugal. O "#" permite indexar a mensagem ao tópico "Portugal" e todos os utilizadores que pesquisarem por este tópico no Twitter irão encontrar tanto esta mensagem, como qualquer outra que utilize o mesmo hashtag. Por outro lado, utilizando o símbolo "@" antes do nome de um utilizador será possível mencionar utilizadores, (como acontece neste exemplo com o utilizador Joaorodrigues77) que receberão uma notificação indicando que foram mencionados num determinado *tweet*. Através dos *tweets* é possível identificar os tópicos indexados e assim, por exemplo, conseguir perceber o que está a ser partilhado no mundo em tempo real.

1.2. *TEXT MINING*

Podemos definir o *Text Mining* como a descoberta, através de recursos informáticos, de informação previamente desconhecida, através da extracção automática de informação proveniente de diferentes fontes de *text*o (Hearst, 2003), de modo a descobrir padrões de informação interessantes e não triviais (Mahesh, Suresh, & Vinayababu, 2010). Pela natureza dos dados das redes sociais (dados não estruturados) torna-se imperativa a criação de novas abordagens e ferramentas para a análise desta informação por forma a obter conhecimento. É neste con*text*o que o *Text Mining* ganha cada vez mais preponderância. Dörre, Gerstl, & Seiffert (1999) identificam o desafio que é obter informação *text*ual de forma computacional tendo em conta a natureza não estruturada destes dados, apontando também diferenças entre Data *Mining* e *Text Mining*. Para os autores, o Data *Mining* envolve de forma geral os seguintes passos:

1. Identificação de uma colecção (dados usados para a modelação)
2. Preparação dos dados e escolha de variáveis
3. Análise de Distribuição

O *Text mining* diferencia-se do Data *Mining* essencialmente nos pontos 2 e 3 (sabendo à partida que no ponto 1 também, uma vez que são dados textuais), uma vez que não é viável em *Text Mining* que o ser humano avalie todas as variáveis a ser ou não incluídas na modelação, tendo em conta que chegam facilmente a números que rondam os milhares, sendo necessária a implementação de algumas diferenças nos algoritmos usados. Temas como matrizes term-by-document e redução de dimensionalidade serão abordados numa fase posterior deste estudo, nas secções 1.5 e 1.7 respectivamente.

Ghosh, Roy, & Bandyopadhyay (2012) identificam os algoritmos de *Text Mining* segundo três categorias. Algoritmos de classificação, descoberta de associações e *Clustering*:

Os algoritmos de **classificação**, consistindo numa base de dados de treino (ou Data Set de treino) composta por variáveis contínuas e discretas, sendo que uma destas variáveis discretas será a nossa "classe". O objetivo será construir um modelo através das restantes variáveis do data set, que permita classificar novos dados, diferentes daqueles que já possuímos. Entre estes encontram-se os modelos de Árvores de Decisão, Redes Neuronais, Algoritmos genéticos, etc.

Podemos também distinguir os algoritmos de **Descoberta de Associações**, que permitem obter conhecimento sobre relações como por exemplo: sendo "A e B implicam C com 80% de certeza". Este tipo de informação poderá ser útil para a tomada de decisão relativa à implementação de determinados aspetos de negócio, que se descubram ser relevantes tendo por base fenómenos observados. Dentro destes, podemos enveredar por algoritmos paralelos e sequenciais para a descoberta de associações.

Poderemos distinguir também os Algoritmos de ***Clustering***. O *Clustering* consiste na divisão dos dados em grupos de indivíduos com características homogéneas entre si (*clusters*) ou intra-*cluster*, e

com características distintivas entre diferentes *clusters*, ou seja inter-*clusters*. Podemos aqui subdividir este tema em algoritmos hierárquicos, e métodos de partição.

Dentro deste con*text*o surge ainda o ***Sentiment Analysis*** (ou Opinion *Mining*). Liu (2012) define-o como a área que analisa as emoções, *sentiment*os e atitudes face a uma determinada entidade, seja ela um produto, serviço, organização, pessoa, eventos, tópicos, etc.

Uma das técnicas utilizadas no âmbito do *Sentiment Analysis* é conhecida como *Bag-of-Words* (BoW), que consiste na análise das palavras (palavra a palavra, sendo cada palavra independente das outras, ou em conjugações de várias palavras ao mesmo tempo) de uma determinada frase ou *text*o com vista à sua classificação de polaridade, ou seja, atribuição de um *sentiment*o positivo, negativo ou neutro do comunicador face ao assunto a que se refere, através da comparação com palavras com sentimentos pré-catalogado (Mudinas, Zhang, & Levene, 2012). Temos também algoritmos mais robustos de machine learning como sendo os classificadores de Bayes, e os Support Vector Machines (SVM) (Mullen & Collier, 2004), que utilizando conjuntos de dados de treino, conseguem "ensinar" o algoritmo a classificar novos exemplos.

Pang, Lee, & Vaithyanathan (2002) utilizaram as técnicas de Naive Bayes, SVM, e Maximum entropy classification (três técnicas de machine learning) com o intuito de classificar críticas de filmes como sendo positivas ou negativas. Concluíram que os modelos de machine learning obtiveram melhores resultados do que as classificações feitas com base num conjunto de palavras escolhidas por dois voluntários humanos como sendo boas para classificar o *sentiment*o das críticas.

Li & Wu (2010) estudaram a detecção de "hotspots" (ou tópicos quentes/actuais/de interesse) em fóruns de desporto online através da classificação de sentimentos nos posts dos fóruns (utilizando uma técnica semelhante ao BoW), *Clustering* de fóruns através do algoritmo *k-means* com vista à identificação dos hotspots e a partir deste, previsão de hotspots baseado num modelo de classificação SVM.

Por outro lado, é comum que os dados sobre os quais as análises anteriormente referidas incidem sejam caracterizados por muitas dimensões. Na realidade, a dimensionalidade é de tal ordem que coloca obstáculos à modelação, levando à maldição da dimensionalidade (Pagolu, hakraborty, 2014). Albright (2004) e Deerswester, Dumais, Furnas, & Landauer (1990) exploram o Singular Value Decomposition (SVD) enquanto técnica de redução de dimensionalidade. Este tema será abordado com maior detalhe na secção 1.7.

METODOLOGIA

Até aqui foi feito um enquadramento do tema. Passaremos a apresentar a organização do estudo. Inicialmente procedeu-se à recolha de dados para o Burger King e o McDonald's. Posteriormente os *tweets* foram pré-processados de modo a serem aplicadas análises sobre os mesmos. Criaram-se matrizes term-by-document para os *tweets* de cada empresa e procedeu-se à aplicação da técnica SVD sobre as matrizes com a finalidade de reduzir a dimensionalidade das mesmas. Posteriormente criaram-se modelos de *Clustering* hierárquico para cada empresa e sobre os *clusters* obtidos aplicou-se *Sentiment analysis* com o intuito de se obter o sentimento associado a *cluster*. O processo descrito encontra-se ilustrado na figura 1.

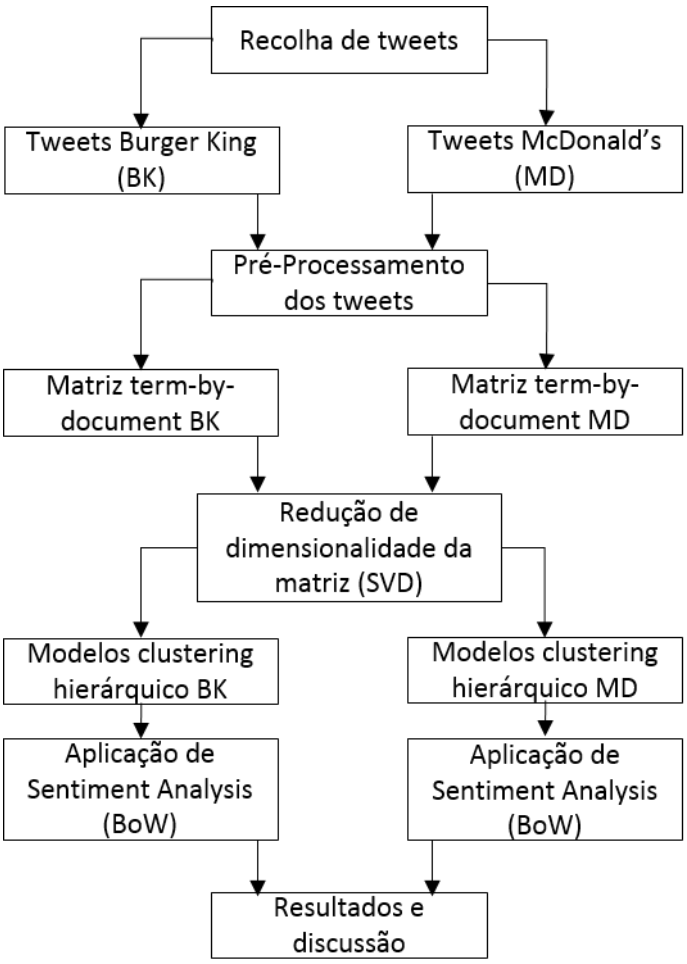

Figura 1 – Diagrama do projecto

Os pontos da metodologia apresentados na figura 1 serão apresentados em seguida de forma mais detalhada.

1.3. RECOLHA DE DADOS

De modo a obter informação para análise foi extraída informação do Twitter (*tweets*) entre 29 de Junho e 20 de Setembro de 2015. Este processo foi possibilitado através da utilização de uma API (Application Programming Interface) do Twitter que permite a ligação aos seus servidores de modo a descarregar informação. A ligação foi estabelecida através de código em linguagem Python (Bernardo & Henriques, 2014). Esta informação foi recolhida 24h por dia durante o período referido sendo um *live feed*, ou seja, extraindo informação à medida que os *tweets* iam sendo feitos pelos utilizadores (com duas horas de atraso face ao momento da publicação).

Para o estudo em causa, foram recolhidos dados de duas empresas que actuam num mercado competitivo: o McDonald's e o Burger King. De modo a procurar informação de *tweets* referente a estas duas empresas, foi usado código Python e termos de pesquisa que consistiram em variações lógicas do nome "oficial" das duas empresas. No caso, os termos de pesquisa usados foram:

Burger King	McDonald's
Burger King	McDonalds
Burger king	Mcdonalds
BurgerKing	macdonalds
burgerking	macDonalds
BURGERKING	mcdonalds
BURGER KING	MCDONALDS
burger king	
Burguer King	
BK	

Tabela 1 - Variantes do nome de cada empresa utilizadas

Com estes termos de pesquisa presentes na tabela 1, para o período de análise em estudo foram recolhidos 1 474 053 *tweets*. Estes distribuem-se da seguinte maneira:

Empresa	Total de *tweets*	Média diária
Burger King	152.843	17.922
McDonald's	1.321.210	28.402

Tabela 2 – Total global e média diária de obtenção de *tweets*

A discrepância verificada no número de *tweets* obtidos para as duas empresas prende-se com o facto de ter existido um erro na definição do nome do Burger King que apenas foi detectado numa fase avançada da extracção de dados. O erro foi corrigido, no entanto tendo em conta a janela temporal definida para a obtenção de informação, obteve-se um número menor de *tweets* face à McDonald's. Adicionalmente, o volume de dados gerados, relativo à empresa McDonald's é muito superior (10.000 *tweets* por dia em média) ao gerado para a empresa Burger King.

1.4. PRÉ-PROCESSAMENTO DOS DADOS

Tendo em conta a natureza dos dados em questão (*texto* não estruturado), é necessário um processamento prévio dos dados em bruto para que sobre estes possam ser aplicadas as análises pretendidas. Para este estudo, o processamento de informação foi feito com recurso a código SAS, através de procedimentos *Perl Regular Expression* (PRX) que são utilizados para reconhecimento de padrões em texto.

- **Caracteres especiais** – foram removidos todos os caracteres especiais do *texto* dos *tweets* (ou seja, tudo o que não fossem letras ou números, exceptuando o símbolo "#" que identifica *hashtags*), uma vez que não são relevantes para a análise, podendo até causar enviesamentos à mesma
- **Menções de outros perfis do** Twitter – o símbolo "@" e correspondente perfil/conta associado foram também removidos do texto uma vez que não atribuem nenhum tipo de significado adicional
- ***Retweets*** – os *retweets* indicam uma partilha de um *tweet* feito por outro utilizador. Apresentam no início do texto a sigla "RT"
- ***Links/URL*** – foram removidos *links/URL* do texto

Antes da criação dos modelos foram removidos *tweets* da base original obtida para as duas empresas. Apenas foram mantidos *tweets* de língua inglesa, foram removidos *tweets* obtidos erroneamente, isto é, que na realidade não correspondiam às empresas em questão, e foram removidos *retweets* uma vez que não representam mais que uma duplicação de um *tweet* original (Choudhary, Singh, & Chakraborty, 2015). São também apenas mantidos os *tweets* que não contando com *hashtags*, tenham pelo menos 4 palavras. Isto deve-se ao facto de um *tweet* com pouco *texto* não ter conteúdo considerado suficientemente significativo para análise (Ifrim, Shi, & Brigadir, 2014).

Após a aplicação destes critérios restou o seguinte número de *tweets*:

Empresa	*Tweets* iniciais	*Tweets* finais	% Retenção
Burger King	152.843	33.098	21,7%
McDonald's	1.321.210	316.181	23,9%

Tabela 3 - Total de *tweets* recolhidos e percentagem de retenção para a modelação

Muitas das palavras decorrentes do uso de linguagem, apesar de serem importantes na articulação da comunicação (seja ela escrita ou falada), não apresentam características descritivas e de significado relevantes, adicionando pouco valor à análise (Gomes, Neto, & Henriques, 2012). Exemplos destas palavras são substantivos, determinantes, preposições, etc. (em inglês alguns exemplos são *"of", "why", "then", "in", "on",* etc.). Estes termos, denominados de "*Stop Words*" incorporam uma lista que foi utilizada para tornar estas palavras "neutras" de modo a impedir a sua influência na modelação. A lista utilizada é a presente por defeito no SAS Enterprise Miner.

Foi também utilizada uma técnica denominada *stemming* que tem como objectivo transformar as palavras na sua forma básica. No caso da língua inglesa, retirar o plural dos nomes, a forma do

gerúndio “-ing” e outros prefixos/sufixos (Hotho, Nürnberger, & PaaB, 2005) . Este processo é interessante pois sem ele teriam de ser reconhecidas pelos algoritmos uma série de variações morfológicas das palavras quando na realidade derivam todas da mesma forma base (Gomes, Neto, & Henriques, 2012). Um exemplo inclui a palavra “*work*” da qual derivam outras como “*working*” ou “*worked*”.

1.5. MATRIZ TERM-BY-DOCUMENT

As matrizes *term-by-document* são um ponto inicial na análise textual. Estas permitem a representação dos *tweets* através de vectores numéricos (*feature vectors*) que possibilitam a modelação deste tipo de dados. O espaço representado pelo conjunto destes vectores é denominado de *Vector Space Model* (VSM). Nestas matrizes bidimensionais as colunas são representados pelos documentos existentes na nossa colecção (no caso todos os nossos *tweets*) e as linhas representam todas as palavras existentes na nossa colecção, ou seja, todas as palavras (distintas) que aparecem no conjunto de todos os *tweets*. Isto dá-nos a frequência de ocorrência para cada palavra existente no nosso universo (Choudhary, Singh, & Chakraborty, 2015). Um exemplo de uma matriz *term-by-document* pode ser visto na tabela 4.

Palavra	Documento 1	Documento 2	Documento 3 ···	Documento n
hungry	2	5	0	1
danger	1	0	0	2
life	0	0	0	1
burguer	0	0	0	0
good	3	1	0	0
love	0	0	1	0
Ice	0	0	0	0
…	…	…	…	…
Palavra n	1	0	9	0

Tabela 4 - Matriz term-by-document

A tabela 4 permite obter a frequência com que cada termo aparece em cada documento (*tweet*). Deste modo, cada *tweet* será representado por um vector de n dimensões. Estas matrizes podem atingir dimensões muito elevadas dependendo do número de documentos e palavras em causa. Pelo facto de serem maioritariamente constituídas por valores zero são denominadas de matrizes dispersas ou esparsas (por oposição a matrizes com maior representatividade de números não nulos, denominadas de matrizes densas). Tal dimensão de dados e consequentes problemas de processamento levam-nos à maldição da dimensionalidade, cuja solução será abordada mais à frente neste estudo na secção 1.7.

1.6. MATRIZ TERM-BY-DOCUMENT PONDERADA

O conteúdo da matriz *term-by-document* por si só representa valores absolutos de frequências. No entanto, pretendemos perceber a importância relativa dos diferentes termos na nossa colecção com vista à compreensão de quais têm de facto relevância face a outros. A ideia é variar a importância dos termos com base na frequência com que os mesmos ocorrem nos seus documentos e como os termos são distribuídos ao longo dos restantes documentos da colecção. Este processo de ponderação pode ser dividido em duas partes. A ponderação local e a ponderação global.

A ponderação local baseia-se na ideia de que palavras que ocorrem frequentemente no con*text*o dos documentos têm algum tipo de preponderância no conteúdo desses textos. Existem algumas funções para o cálculo desta medida. A forma básica ou *raw frequency* é calculada como sendo o número de ocorrências dessa palavra nesse documento específico. No entanto, neste estudo é utilizada uma fórmula logarítmica para a frequência (Sanders & DeVault, 2004):

$$Log_2\left(1 + F_{t,d}\ \right) \tag{3.1}$$

Onde $F_{t,d}$ representa a frequência da palavra *t* no documento *d*. A utilização desta função suaviza o efeito de palavras que aparecem muitas vezes num documento. No entanto, a utilização isolada desta estatística levaria a que palavras que ocorrem muitas vezes tivessem inevitavelmente uma maior preponderância. É aqui que a ponderação global se torna útil. A ponderação global pode ser utilizado para identificar a importância de cada termo face a uma colecção e assignar-lhes pesos (Murugesan & Zhang, 2011). Apresentamos aqui dois métodos.

O ***Inverse Document Frequency (IDF)*** é uma medida de ponderação global que tem como objectivo medir a "raridade" de um termo na totalidade da colecção (Roul, Kumar, Devanand, & Sahay, 2014). Ou seja, esta estatística será tanto maior quanto menor for o número de documentos em que a palavra aparece. A fórmula de cálculo é a seguinte:

$$log_2\left(\frac{1}{P_{(t_i)}}\right) + 1 \tag{3.2}$$

Onde $P(t_i)$ representa a proporção de documentos que têm a palavra t_i.

Outra técnica denominada *Entropy*, à semelhança do IDF, dá também um maior peso aos termos que aparecem com menor frequência na colecção ao usar um derivado da medida de entropia (Dong, 2008).

$$1 + \sum_j \frac{(f_{ij}/g_i).log_2(f_{ij}/g_i)}{log_2(n)} \tag{3.3}$$

Onde g_i é o número de vezes que o termo *i* aparece na colecção e n é o número de documentos na colecção.

Cada palavra da matriz *term-by-document* terá então uma frequência ponderada associada dada pelo produto da ponderação local e da ponderação global, sendo que a ponderação global poderá ser IDF (LOG-IDF) ou *Entropy* (LOG-*Entropy*). Os pesos locais e globais tentam equilibrar a noção de que uma palavra que apareça muitas vezes ao longo de toda a colecção tende a não ser discriminante, mas se aparecer em poucos documentos e tiver frequência elevada em algum documento, então a palavra será discriminante (Dong, 2008). Um resultado final elevado para o ponderador será então obtido se o valor local for alto (se o termo ocorrer muito frequentemente em algum documento) e se o termo tiver uma frequência baixa ao longo da colecção.

Será utilizado a ponderação local LOG em conjugação com a ponderação global *Entropy*, sendo esta a configuração pré-definida do SAS Enterprise Miner. Para além desta, será também utilizada a conjugação LOG-IDF.

1.7. SINGULAR VALUE DECOMPOSITION (SVD)

O *Singular Value Decomposition* (SVD) é uma técnica de redução de dimensionalidade que permite converter a matriz dispersa de frequências term-by-document numa matriz densa e portanto de dimensão mais reduzida, que facilita o processamento de dados. A ideia será passar de um espaço dimensional n para um espaço dimensional k sendo k<r (r representa a característica da matriz A). Suponhamos que temos uma matriz dispersa A. O SVD desta matriz será a decomposição da matriz original A em 3 novas matrizes (Pagolu, Murali, Chakraborty, & Goutam, 2004):

$$A = U\Sigma V^T \tag{3.4}$$

Em que:

U é uma matriz ortogonal cujas colunas são chamadas de "*left singular vectors*"
Σ é uma matriz diagonal cujos elementos da diagonal principal são chamados de "*singular values*"
V é uma matriz ortogonal cujas colunas são chamadas de "*right singular vectors*"

A aproximação da matriz A reduzida é dada por:

$$A_k = U_k \Sigma_k {V_k}^T \tag{3.5}$$

Sendo *k* a característica da nova matriz reduzida.

O cálculo SVD da matriz A está intimamente ligado à decomposição em *eigenvectors*/*eigenvalues*. O objectivo da aplicação do SVD é a redução de dimensionalidade, ou seja, colocar os documentos (ou termos) num espaço dimensional menor que o original. Assumindo as colunas da matriz A como coordenadas de um espaço multidimensional, cada coluna representa um ponto (documento, ou no nosso caso um *tweet*) nesse espaço. Do mesmo modo, assumindo como exemplo a primeira coluna da matriz U, esta representa um ponto no espaço. Esta primeira coluna de U minimiza a distância às colunas originais de A (numa óptica de mínimos quadrados). A segunda coluna de U em conjunto com a primeira criam um plano que é um subespaço bi-dimensional do original que minimiza novamente a distância, e assim sucessivamente. Assim sendo, as primeiras *k* colunas de U representam um subespaço do original que é o melhor ajustado em termos de mínimos quadrados (SAS Institute Inc., 2001). De forma genérica estas colunas podem ser vistas como conceitos "artificiais", isto é, significado extraído dos diferentes documentos/palavras (Deerswester, Dumais, Furnas, & Landauer, 1990). Podemos então projectar as colunas de A nas *k* primeiras colunas de U para reduzir a dimensionalidade. Cada documento *d*, representado por um vector de palavras, será projectado neste novo e mais reduzido espaço sub-dimensional através da seguinte fórmula:

$$\hat{d} = {U^T}_k d \tag{3.6}$$

Esta matriz U (composta pelos *left singular vectors*) permite-nos em conjunto com a matriz *term-by-document*, obter as projecções SVD. O *k* é definido *a priori* pelo utilizador, devendo *k* ser um número suficientemente grande para apanhar os padrões existentes nos dados, mas não tão grande que apanhe "ruído". Albright (2004) exemplifica o cálculo do SVD para *k*=2:

	Documentos		
Termo	d1	d2	d3
1. Error:	1	1	1
2. invalid	1	0	0
3. message	1	2	0
4. file	1	1	0
5. format	1	0	1
6. unable	0	1	1
7. to	0	1	1
8. open	0	1	0
9. using	0	1	0
10. path	0	1	0
11.variable	0	0	1

Tabela 5 - Exemplo de tabela de frequências term-by-document

A partir da matriz *term-by-document* de exemplo apresentada na tabela 5 é possível obter as restantes matrizes:

U2	
0,43	0,3
0,11	0,13
0,55	-0,37
0,33	-0,12
0,21	0,55
0,31	0,18
0,31	0,18
0,22	-0,25
0,22	-0,25
0,22	-0,25
0,09	0,42

Tabela 6 – Matriz U de *left singular vectors* para *k*=2

Σ2	
3.79	0
0	1.96

Tabela 7 – Matriz Σ de *singular values* para *k*=2

V2	
0,43	0,25
0,82	-0,49
0,36	0,83

Tabela 8 – Matriz V de singular values para *k*=2

Projectando o documento 1 que é formado pelas palavras *Error:, invalid, message, file e format* num espaço bi-dimensional, obtemos:

$$d = [1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0]^T$$

Aplicando a projecção SVD obtemos:

$$\hat{d} = U^T{}_2 d = [1.63, .49]^T$$

O documento 1 passa a ser representado da seguinte forma:

d1	
SVD1	1,63
SVD2	0,49

Tabela 9 – Matriz com valores SVD

Esta transformação leva a uma redução de 11 variáveis originais para 2. As 2 novas variáveis são combinações lineares das 11 variáveis originais, sendo obtidas através dos pesos provenientes da matriz U. Cada documento ou palavra passa então a ser caracterizado por um vector de pesos que indicam a força de associação a cada um dos conceitos (*k* colunas de U) anteriormente referidos, isto é, o "significado" de um documento original pode ser expresso por *k* factores (Albright, 2004).

1.8. *CLUSTERING*

O *Clustering* baseia-se no cálculo de distâncias entre observações sendo utilizado para segmentar informação (Mosley Jr & Roosevelt, 2012). O objectivo é obter *clusters* (também denominados de grupos ou segmentos) que são semelhantes entre si (intra-*cluster*) e diferentes dos restantes (inter-*cluster*). No con*exto* de *text Mining*, os algoritmos de *Clustering* dividem a nossa colecção de documentos em grupos mutuamente exclusivos com o intuito de identificar a presença de temas semelhantes (Pagolu, Murali, Chakraborty, & Goutam, 2004). A ideia por trás do *Clustering* de *tweets* será a de que *tweets* pertencendo a um mesmo tópico irão juntar-se, pelo que poderemos considerar um tópico detectado (Ifrim, Shi, & Brigadir, 2014). Tendo em conta o volume de dados gerado pelas plataformas socias, no caso o Twitter, não é exequível o tratamento manual deste tipo de dados, pelo que os algoritmos de *Clustering* surgem como uma potencial solução para obter *insight* sobre os clientes da empresa. Tendo em conta o exposto, foram criados modelos de *Clustering* com o intuito de segmentar a base de *tweets* para o Burger King e para a McDonald's. Foi utilizado o SAS Enterprise Miner para a criação dos modelos tendo sido utilizado o algoritmo de *clustering* hierárquico.

1.8.1. *Clustering* Hierárquico

O *Clustering* hierárquico é uma técnica que consiste na construção de uma hierarquia de segmentos. O processo pode ser desenvolvido de dois modos (Maimon & Lior, 2005):

Divisiva (*Top-Down*) - As observações começam em um único *cluster* e vão sendo divididas até que seja atingido um critério de paragem.

Aglomerativa (*Bottom-Up*) - Cada observação começa no seu próprio *cluster* e os *clusters* que se encontram mais próximos uns dos outros vão sendo fundidos. O processo itera até que seja atingido um critério de paragem.

De modo a fazer as agregações/divisões de *clusters* é necessária a noção de distância para definir as proximidades. Modo geral, os métodos de *Clustering* definem uma medida de distância entre observações numa base de dados/repositório de informação. Estas observações são agrupadas por forma a minimizar as distâncias entre os membros de cada grupo (Pedersen & Bruce, 1997).

Neste estudo foi utilizado o método *Ward's Minimum Variance* que está incluído na categoria de *clustering* aglomerativo. Segundo Pedersen & Bruce (1997) começamos com N *clusters* e posteriormente os 2 *clusters* mais próximos são agregados dando origem a um novo *cluster*, sendo aqui os *clusters* representados por documentos (*tweets*). Este processo é repetido até ser atingido um número de *clusters* especificado *a priori,* representando este número a quantidade de "significados" (temas) que pretendemos distinguir. Neste método é calculada a variância interna de cada *cluster*, dada pelo somatório do quadrado das diferenças entre cada observação e a média desse *cluster* (isto é, a média dos *feature vectors* nesse *cluster*). A cada iteração, na fusão, é criado um *cluster* com a menor variância interna possível, sendo criado a partir dos dois *clusters* C_K e C_L com a menor variância entre si. O cálculo da variância entre os dois *clusters* C_K e C_L é dado por:

$$V_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{\frac{1}{N_K} + \frac{1}{N_L}} \quad (3.7)$$

Onde $\bar{x}_K$ representa a média do *cluster* K e N_K representa o total de observações do *cluster K*. O mesmo se aplica ao *cluster L*. O numerador da equação representa o quadrado da distância euclidiana entre as médias dos 2 *clusters*.

Foi utilizada a medida *Root Mean Square Standard Deviation* (RMSSTD) para avaliar a qualidade dos modelos. Esta medida é definida como a raíz quadrada da variância de todas as variáveis que formam o *cluster* (Halkidi, Batistakis, & Vazirgiannis, 2002). Como o objectivo do *Clustering* é criar grupos homogéneos, esta medida deverá ser o mais pequena possível. Quanto menor o RMSSTD mais homogéneo é o *cluster*.

Não existe um método que determine o número perfeito de *clusters* que deverão ser gerados pelo algoritmo (Carnegie Mellon University, 2009). O SAS Enterprise Miner permite a definição do número de *clusters* a ser gerado pelo algoritmo. O valor pré-definido é de 40 *clusters*. O algoritmo foi corrido inicialmente com este valor pré-definido e posteriormente com valores que variaram entre 10 e 40. Este intervalo de *clusters* gerou muitos segmentos o que tornou confusa a interpretação dos mesmos, tendo sido definido que o algoritmo deveria gerar um máximo de 10 *clusters* (ou seja, itera até um máximo de 10 *clusters* podendo o número final obtido ser inferior ou igual a 10). Este número foi escolhido para que se conseguissem extrair suficientes "temas".

Após a formação dos *clusters* são apresentadas as palavras que melhor definem esses mesmos *clusters*. As palavras são escolhidas através da função distribuição cumulativa binomial dada por:

$$Prob = F(k|N, p) \quad (3.8)$$

Onde:

k - ocorrências da palavra no *cluster* j

N – número de documentos no *cluster* j

p – (ocorrências da palavra em todos os documentos – *k*)/(Total de documentos - *N*)

As palavras que irão representar o *cluster* serão as que tiverem as probabilidades binomiais mais elevadas, sendo os termos apresentados por ordem de probabilidade descendente (Sanders & DeVault, 2004). No caso deste estudo, foram escolhidas 6 palavras para representar o segmento, pelo que serão as 6 palavras com as probabilidades binomiais mais elevadas. Foram escolhidos 6 termos para que fosse mantida a capacidade de interpretabilidade dos temas dos *clusters* sem adicionar muito "ruído", uma vez que sendo apresentadas por ordem decrescente de importância são estas as mais representativas do *cluster*.

Os modelos de *Clustering* hierárquico foram criados tendo por base algumas definições, que deram origem a modelos distintos, nomeadamente:

- Número de dimensões SVD utilizadas
- Utilização ou não de *stemming* de palavras
- Conjugação dos métodos de ponderação de termos que vimos anteriormente.

1.9. *SENTIMENT ANALYSIS*

Os *tweets* presentes na nossa coleção foram submetidos à análise de *sentiment*os (ou *opinion mining*). Este procedimento tem como objectivo final a extracção da polaridade inerente a cada *tweet*, ou seja, classificar os *tweets* com um de três possíveis outputs: positivo, negativo ou neutro. Tendo em vista este objectivo será utilizada uma técnica intitulada *Bag-of-Words*.

1.9.1. *Bag-of-Words* (BoW)

Na abordagem *Bag-of-Words* (BoW) cada documento é representado como um vector de palavras que ocorrem nesse documento (Matsubara, Takashi, Martins, & Monard, 2003). No caso deste projecto os documentos representam os *tweets*. Querendo isto dizer que cada *tweet* passará a ser representado por um vector, sendo esse vector constituído por um número para cada palavra existente no *tweet*. A título de exemplo, consideremos os seguintes documentos:

Tweet 1: "*I love those burguers at Burger King*!"

Tweet 2: "*McDonald's has the best burguers!! I want them really bad!*"

Com base na frequência de ocorrência de palavras existentes nos dois *tweets* apresentados anteriormente, podemos obter os mesmos dois *tweets* representados por vectores numéricos. Para efeitos de análise de *sentiment*os podemos utilizar esta representação vectorial através dos valores "1", "0" e "-1", sendo o valor "1" atribuído a uma palavra com conotação positiva, o valor "-1" atribuído a uma palavra negativa e "0" para uma palavra neutra. No caso do *tweet* 1 ficaria:

Tweet 1 : ["I", "love", "those", "burguers", "at", "Burger", "King"] → [0,1,0,0,0,0,0]

O *tweet* é então classificado através do score obtido, ou seja, se o score for superior a zero será positivo, se o score for inferior a zero será negativo, e se for igual a zero será neutro. No caso exemplificado, o *tweet* 1 obtém um score final de "1", ou seja, positivo. No caso do *tweet* 2 ficaria:

tweet 2 : ["McDonald's", "has", "the", "best", "burguers", "I", "want", "them", "really", "bad"] →

→ [0,0,0,1,0,0,0,0,0,-1]

No caso exemplificado, o *tweet* 2 obtém um score final de "0", ou seja neutro, uma vez que neste caso temos uma palavra positiva e outra negativa, o que nos dá um resultado neutro. A fórmula aplicada no cálculo do *sentiment*o a atribuir é a seguinte:

$$\sum_{i=1}^{n} Palavra\ i \qquad (3.9)$$

Onde i=índice da palavra no *tweet* (palavra 1, palavra 2, etc.) e n=total de palavras do *tweet*.

Mas como consegue o algoritmo perceber se uma palavra tem conotação positiva ou negativa? Hu & Liu (2004) utilizam o BoW para obter um vector de palavras pré-classificadas (a que podemos chamar de dicionário) com o qual as palavras dos nossos *tweets* são comparadas de modo a que se possa obter a polaridade.

Existem algumas limitações neste método. Nomeadamente:

"*I wish this burguer tasted <u>better</u>*!"

O *tweet* exemplificado seria classificado como positivo uma vez que que a palavra "better" tem uma conotação positiva. No entanto, no con*text*o da frase, o *tweet* na realidade é negativo, pois o utilizador está a demonstrar desagrado com o hambúrguer que consumiu. O BoW não consegue portanto distinguir palavras dentro de um con*text*o, isto é, limita-se às palavras em si.

O algoritmo foi corrido através de um *script* em linguagem R (Miner, 2012) para a atribuição do sentimento.

RESULTADOS E DISCUSSÃO

1.10. MODELOS DE *CLUSTERING*

Com recurso ao SAS Enterprise Miner e tendo como base de partida o número de *clusters* obtidos na tabela 3, foram então criados os modelos de *Clustering* Hierárquico para o Burger King e para o McDonald's.

1.10.1. Modelos Burger King

MODELO 1					
#	*Clusters* obtidos	Nome	Frequência	% do total	RMS STD
1	'mcwhopper burger king' 'mcwhopper saga' +sandwich +team epic mcwhopper	Mcwhopper	3283	9,92%	0,102
2	'burger king trap' +hen +house +king +peace +trap	Música "Burger King Trap House"	18295	55,28%	0,117
3	+airport +big +cream +hungry +ice +shake	Produtos 1 (ice, cream, shake)	7194	21,74%	0,111
4	'chicken fries' +chicken +good +time arrested best	Produtos 2 (chicken)	4326	13,07%	0,110

Tabela 10 – Modelo 1 de *Clustering* Hierárquico para o Burger King

O primeiro modelo de *Clustering* criado para o Burger King obteve 4 *clusters*, como ilustrado na tabela 10. O modelo em questão foi criado com 100 dimensões SVD, *stemming* de palavras e com a ponderação LOG-IDF. Sendo um modelo que identifica temáticas fortemente ligadas a produtos. Em média o RMS STD do modelo é de 0,11.

1. Mcwhopper - O primeiro *cluster* identifica o tema "Mcwhopper". Mcwhopper representa a proposta de criação de um hambúrguer conjunto que o Burger King fez ao McDonald's na tentativa de fazer "tréguas" com a empresa rival como forma de celebração do dia Internacional da Paz, comemorado a 21 de Setembro. Proposta que o McDonald's rejeitou.

2. Música "Burger King Trap House" – Este *cluster* identifica a música "Burger King Trap House" criada por um artista norte-americano chamado Jay Hen Gwoppa.

3. Produtos 1 (ice, cream, shake) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente gelados e batidos.

4. Produtos 2 (chicken) - O *cluster* identifica produtos comercializados pela empresa, nomeadamente frango.

MODELO 2					
#	*Clusters* obtidos	Nome	Frequência	% do *cluster*	RMS STD
1	'peace day' +'peace burger' +'peace day burger' +create +offer +peace	Mcwhopper	1180	3,57%	0,083
2	'burger king trap' +house +sandwich +trap +want epic	Música Burger King Trap House	5522	16,68%	0,079
3	'donald trump' +good +life +love +meal +stop	?	5041	15,23%	0,072
4	+breakfast +chicken +king +right +team +time	Serviços/Produtos (breakfast,chicken)	21355	64,52%	0,084

Tabela 11 – Modelo 2 de *Clustering* Hierárquico para o Burger King

O segundo modelo para o Burger King obteve 4 *clusters*, como ilustrado na tabela 11. O modelo em questão foi criado com 135 dimensões SVD, *stemming* de palavras e com a ponderação LOG-IDF. Sendo um modelo que identifica temáticas ligadas a produtos e serviços. Em média o RMS STD do modelo é de 0,078.

1. Mcwhopper - O primeiro *cluster* identifica o tema "Mcwhopper". Mcwhopper representa a proposta de criação de um hambúrguer conjunto que o Burger King fez ao McDonald's na tentativa de fazer "tréguas" com a empresa rival como forma de celebração do dia Internacional da Paz, comemorado a 21 de Setembro. Proposta que o McDonald's rejeitou.

2. Música "Burger King Trap House" – Este *cluster* identifica a música "Burger King Trap House" criada por um artista norte-americano chamado Jay Hen Gwoppa.

3. ? – Este *cluster* não permite obter uma noção bem definida das temáticas abordadas.

4. Serviços/Produtos (breakfast,chicken) - O *cluster* identifica produtos comercializados pela empresa, nomeadamente frango, e também serviços, no caso os pedidos feitos nas lojas.

MODELO 3					
#	*Clusters* obtidos	Nome	Frequência	% do *cluster*	RMS STD
1	'donald trump' 'florida man' arrested donald florida last	?	4906	14,82%	0,109
2	always craving first good im	Sensações	8615	26,03%	0,108
3	'burger king trap' chicken fries getitlive house mcwhopper	Música Burger King Trap House/produtos	12118	36,61%	0,117
4	'burger king trap' gwoppa hen house jay king	Música Burger King Trap House	7459	22,54%	0,117

Tabela 12 – Modelo 3 de *Clustering* Hierárquico para o Burger King

O segundo modelo para o Burger King obteve 4 *clusters*, como ilustrado na tabela 12. O modelo em questão foi criado com 100 dimensões SVD, sem *stemming* de palavras e com a ponderação LOG-Entropy. Sendo um modelo que identifica temáticas ligadas a produtos e serviços. Em média o RMS STD do modelo é de 0,113.

1. ? - Este *cluster* não permite obter uma noção bem definida das temáticas abordadas.

2. Sensações – Este *cluster* identifica sensações ("*cravings*", ou em português desejos/ânsias).

3. Música Burger King Trap House/produtos – Este *cluster* identifica a música "Burger King Trap House" criada por um artista norte-americano chamado Jay Hen Gwoppa, bem como alguns produtos (frango e batatas fritas).

4. Música Burger King Trap House - Este *cluster* identifica a música "Burger King Trap House" criada por um artista norte-americano chamado Jay Hen Gwoppa.

1.10.2. Modelos McDonald's

MODELO 1					
#	*Clusters* obtidos	Nome	Frequência	% do *cluster*	RMS STD
1	'mcdonalds fries' +'mcdonalds breakfast' +'sweet tea' +breakfast +drive +job	Produtos 1 (Fries, breakfast, tea)	33318	10,54%	0,106
2	+dinner +fat +feel +hash +sick doesn	Sensações	22312	7,06%	0,104
3	+'happy meal' +car +coffee +cream +down +free	Produtos 2 (Happy meal, coffee, cream)	62771	19,85%	0,113
4	chicken nuggets' +big +chicken +good +money	Produtos 3 (Chicken nuggets)	82725	26,16%	0,107
5	+last +lemonade +order +stop +strawberry +want	Produtos 4 (lemonade, strawberry), Serviço (Order)	60301	19,07%	0,112
6	'french fries' +find +great +nice +start +wish	Produtos 5 (French fires)	54754	17,32%	0,099

Tabela 13 – Modelo 1 de *Clustering* Hierárquico para o McDonald's

O primeiro modelo de *Clustering* criado para o McDonald's obteve 6 *clusters*, como ilustrado na tabela 13. O modelo em questão foi criado com 100 dimensões SVD, *stemming* de palavras e com a ponderação LOG-IDF. Sendo um modelo que identifica temáticas fortemente ligadas a produtos. Em média o RMS STD do modelo é de 0,107.

1. Produtos 1 (Fries, breakfast, tea) - O *cluster* identifica produtos comercializados pela empresa, nomeadamente batatas fritas e chá, identificando também os pequenos-almoços.

2. Sensações – O *cluster* identifica *sentimen*tos/sensações parecendo fazer referência a *sentimen*tos de mal-estar (fat, feel, sick) que poderão indiciar malefícios para a saúde da ingestão decorrente dos produtos da empresa.

3. Produtos 2 (Happy meal, coffee, cream) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente o Happy Meal, o café e as natas.

4. Produtos 3 (Chicken nuggets) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente os nuggets de frango.

5. Produtos 4 (lemonade, straberry), Serviço (Order) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente a limonada e morangos, bem como o serviço de pedidos nas lojas

6. Produtos 5 (French fires) - O *cluster* identifica produtos comercializados pela empresa, nomeadamente as batatas fritas.

MODELO 2					
#	*Clusters* obtidos	Nome	Frequência	% do *cluster*	RMS STD
1	free gonna good job menu minions	Job/minions	55585	17,58%	0,112
2	best friend gonna hour hungry im	?	18999	6,01%	0,104
3	'big mac' big happy hour mac meal	Produtos 1 (Big Mac, Happy Meal)	14495	4,58%	0,108
4	'first time' 'happy meal' 'mcdonalds today' eating first happy	Produtos 2 (Happy Meal), Experiência (primeira vez no McDonald's)	22472	7,11%	0,107
5	coffee doesn hours house iced late	Produtos 3 (coffee)	23406	7,40%	0,099
6	bad hungry im rn want dont	?	14970	4,73%	0,087
7	better chicken cream drive ice	Produtos 4 (chicken, cream)	98140	31,04%	0,109
8	days dinner great having kfc last	KFC	44285	14,01%	0,093
9	'french fries' 'mcdonalds breakfast' 'mcdonalds fries' breakfast french fries	Produtos 5 (French fries, breakfast)	23829	7,54%	0,106

Tabela 14 – Modelo 2 de *Clustering* Hierárquico para o McDonald's

O segundo modelo de *Clustering* criado para o McDonald's obteve 9 *clusters*, como ilustrado na tabela 14. O modelo em questão foi criado com 100 dimensões SVD, sem *stemming* de palavras e com a ponderação LOG-IDF. É um modelo que aparenta não identificar temas de forma muito clara/coerente. Em média o RMS STD do modelo é de 0,103.

1. Job/minions – O *cluster* identifica a oportunidade de trabalhar no McDonald's bem como os artigos de merchandising do filme "Minions" presentes nos produtos do McDonald's.

2. ? - Este *cluster* não permite obter uma noção bem definida das temáticas abordadas.

3. Produtos 1 (Big Mac, Happy Meal) – O *cluster* identifica produtos comercializados pela empresa, o Big Mac e o Happy Meal.

4. Produtos 2 (Happy Meal), Experiência (primeira vez no McDonald's) – O *cluster* identifica produtos comercializados pela empresa e aborda a experiência de experimentar o McDonald's pela primeira vez.

5. Produtos 3 (coffee) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente café.

6. ? - Este *cluster* não permite obter uma noção bem definida das temáticas abordadas.

7. Produtos 4 (chicken, cream) - O *cluster* identifica produtos comercializados pela empresa, nomeadamente frango e natas.

8. KFC – O *cluster* aparenta fazer referência a um concorrente (o KFC)

9. Produtos 5 (French fries, breakfast) - O *cluster* identifica produtos comercializados pela empresa, nomeadamente batatas fritas e os pequenos-almoços.

MODELO 3					
#	*Clusters* obtidos	Nome	Frequência	% do *cluster*	RMS STD
1	best chicken cream eating free ice	Produtos 1 (Chicken, Ice cream)	88620	28,03%	0,106
2	'big mac' 'mcdonalds breakfast' apple big breakfast coke	Produtos 2 (Big Mac, coke)	30122	9,53%	0,104
3	hard hope interview job machine miss	Job/Interview	12727	4,03%	0,097
4	'french fries' 'mcdonalds fries' chicken damn french fries	French Fries	31599	9,99%	0,105
5	'parking lot' drive good last lot love	Serviço (McDrive)	23556	7,45%	0,106
6	better coffee hate im lol	?	101840	32,21%	0,109
7	'first time' 'mcdonalds today' bought didn feel first	Experiência (primeira vez no McDonald's)	27717	8,77%	0,107

Tabela 15 – Modelo 3 de *Clustering* Hierárquico para o McDonald's

O terceiro modelo de *Clustering* criado para o McDonald's obteve 9 *clusters*, como ilustrado na tabela 15. O modelo em questão foi criado com 100 dimensões SVD, sem *stemming* de palavras e com a ponderação LOG-Entropy. É um modelo que aparenta não identificar temas de forma muito clara/coerente. Em média o RMS STD do modelo é de 0,119.

1. Produtos 1 (Chicken, Ice cream) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente frango e gelados.

2. Produtos 2 (Big Mac, coke) – O *cluster* identifica produtos comercializados pela empresa, nomeadamente o Big Mac e a Coca-Cola.

3. Job/Interview – O *cluster* identifica oportunidades de trabalhar no McDonald's.

4. French Fries – O *cluster* identifica produtos comercializados pela empresa, nomeadamente battas fritas.

5. McDrive – O *cluster* identifica um serviço (McDrive).

6. ? - Este *cluster* não permite obter uma noção bem definida das temáticas abordadas.

7. Experiência (primeira vez no McDonald's) - O *cluster* aborda a experiência de experimentar o McDonald's pela primeira vez.

1.11. MODELOS DE *SENTIMENT ANALYSIS*

Para o desenvolvimento dos modelos de análise de *sentimento*s foram classificados todos os *tweets* da base de *tweets* do Burger King, e todos os *tweets* da base de *tweets* do McDonald's. A distribuição do *sentimento* para as duas empresas é a seguinte:

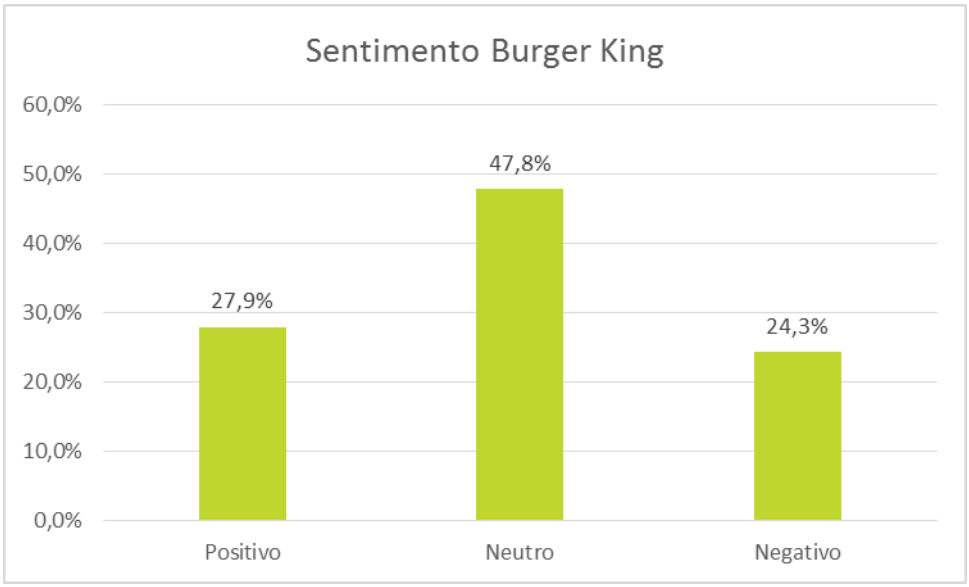

Figura 2 – Distribuição de *sentimento* para o Burger King

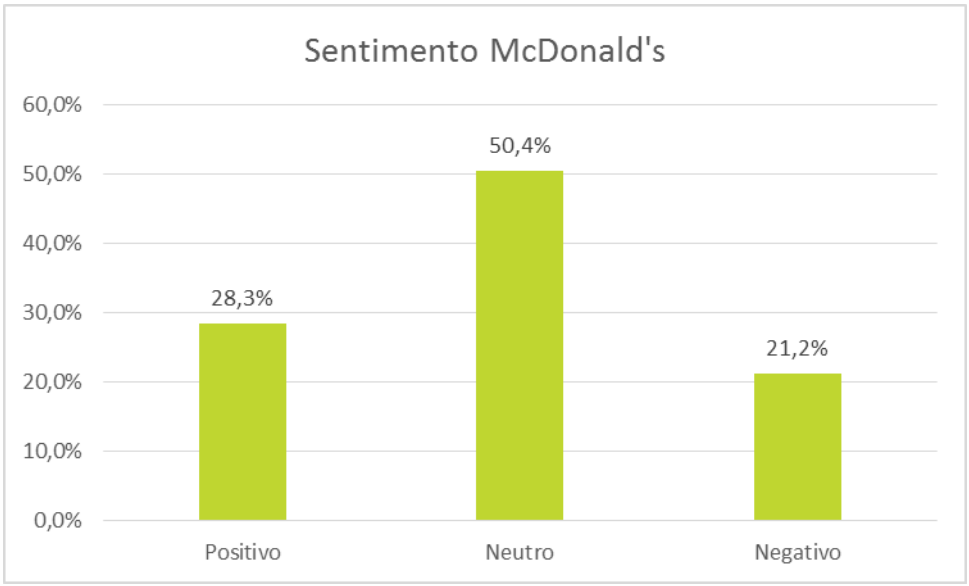

Figura 3 - Distribuição de *sentimento* para o McDonald's

Posteriormente, decorrente desta classificação, o *sentimento* foi alocado aos respectivos *tweets* de cada modelo obtido no *Clustering* hierárquico, para as duas empresas.

Foi criada uma escala de polaridade para classificar os *clusters* tendo em conta a proporção de *tweets* positivos face aos negativos, sendo que os neutros, apesar de serem contabilizados, não

interferem directamente na atribuição do *sentimento* geral do *cluster*. A escala atribui a polaridade através da diferença em pontos percentuais entre a percentagem de *tweets* positivos e a percentagem de *tweets* negativos. Por exemplo, se a percentagem de *tweets* positivos de um segmento for de 25% e a de *tweets* negativos for de 15%, a diferença *d* será de 25%-15% = 10 pontos percentuais, o que segundo a escala apresentada na tabela 16 atribuiria um sentimento positivo ao *cluster*.

Diferença de *tweets* Positivos face aos Negativos (*d*)	Classificação
d >= 15	Muito Positivo
0 < *d* < 15	Positivo
d = 0 (com arredondamento às unidades)	Neutro
-15<= *d* < 0	Negativo
d < -15	Muito Negativo

Tabela 16 – Escala de polaridade

Aplicando esta escala aos modelos de *Clustering* obtidos anteriormente, obtemos o sentimento para cada *cluster*.

1.11.1. Modelos Burger King

MODELO 1					
Cluster	*Sentimento*	Frequência	Total de *tweets* do *cluster*	Percentagem	*Sentimento* do *Cluster*
Mcwhopper	Negativo	625	3283	19,0%	Neutro
	Neutro	2026		61,7%	
	Positivo	632		19,3%	
Música “Burger King Trap House”	Negativo	4692	18295	25,6%	Positivo
	Neutro	8318		45,4%	
	Positivo	5293		28,9%	
Produtos 1 (ice cream, shake)	Negativo	1884	7194	26,2%	Negativo
	Neutro	3745		52,1%	
	Positivo	1565		21,8%	
Produtos 2 (chicken)	Negativo	840	4326	19,4%	Muito Positivo
	Neutro	1728		39,9%	
	Positivo	1760		40,7%	

Tabela 17 – *Sentiment Analysis* para os *clusters* do modelo 1 do Burger King

Mcwhopper - O *cluster* que identifica o tema McWhopper é classificado como neutro (19%-19%) dando a entender que a polémica da criação de um hambúrguer conjunto entre as duas empresas rivais dividiu a audiência.

Música “Burger King Trap House” – O segmento com a música foi marcado como tendo um *sentimento* global positivo (29%-26%).

Produtos 1 (ice cream, shake) – O *cluster* Produtos 1 que identifica gelados e batidos, tem um *sentimento* negativo associado (22%-26%). O que poderá indiciar algum problema identificado pelos *clusters* na qualidade destes produtos ou na infraestrutura que os suporta.

Produtos 2 (chicken) - Por outro lado o segmento que identifica frango tem um *sentimento* muito positivo associado (41%-19%), o que indicará que são produtos com boa cotação entre os clientes.

MODELO 2					
Cluster	***Sentimento***	**Frequência**	**Total de *tweets* do *cluster***	**Percentagem**	***Sentimento* do *Cluster***
Mcwhopper	Negativo	736	5041	14,6%	Muito Positivo
	Neutro	2005		39,8%	
	Positivo	2300		45,6%	
Música Burger King Trap House	Negativo	2437	5522	44,1%	Muito Negativo
	Neutro	1813		32,8%	
	Positivo	1272		23,0%	
?	Negativo	32	1180	2,7%	Muito Positivo
	Neutro	217		18,4%	
	Positivo	931		78,9%	
Serviços/Produtos (breakfast,chicken)	Negativo	4836	21355	22,6%	Negativo
	Neutro	11782		55,1%	
	Positivo	4747		22,2%	

Tabela 18 –*Sentiment Analysis* para os *clusters* do modelo 2 para o Burger King

Mcwhopper – Neste modelo o *cluster* que identifica o tema McWhopper é classificado como muito positivo (46%-15%) dando a entender que a polémica da criação de um hambúrguer conjunto entre as duas empresas impactou positivamente na audiência.

Música "Burger King Trap House" – O segmento com a música foi marcado como tendo um *sentimento* global muito negativo (23%-44%), por oposição ao modelo anterior.

? – Acrescentando ao facto de não ser possível discernir um tema ou conjunto de temas deste *cluster*, é classificado como Muito positivo com uma proporção muito desigual de positivos e negativos (79%-3%).

Serviços/Produtos (breakfast,chicken) – O *cluster* identifica serviços/produtos no caso o pequeno-almoço e os produtos de frango da cadeia. Tem um *sentimento* negativo associado (22,2%-22,6%) apesar de existir pouca diferença. É também o maior *cluster* do modelo com 65% do total de *tweets*.

MODELO 3					
Cluster	***Sentimento***	**Frequência**	**Total de *tweets* do *cluster***	**Percentagem**	***Sentimento* do *Cluster***
?	Negativo	784	4906	15,98%	Muito Positivo
	Neutro	2279		46,45%	
	Positivo	1843		37,57%	
Sensações	Negativo	2592	8615	30,09%	Negativo
	Neutro	4076		47,31%	
	Positivo	1947		22,60%	
Música Burger King Trap House/produtos	Negativo	2595	12118	21,41%	Positivo
	Neutro	5702		47,05%	
	Positivo	3821		31,53%	
Música Burger King Trap House	Negativo	2068	7459	27,72%	Negativo
	Neutro	3756		50,36%	
	Positivo	1635		21,92%	

Tabela 19 –*Sentiment Analysis* para os *clusters* do modelo 3 para o Burger King

? – Acrescentando ao facto de não ser possível discernir um tema ou conjunto de temas deste *cluster*, é classificado como Muito positivo com uma proporção muito desigual de positivos e negativos (38%-16%).

Sensações – O *cluster* que identifica sentimentos/sensações apresenta um sentimento global negativo (23%-30%).

Música Burger King Trap House/produtos – O segmento com a música e alguns produtos comercializados pela empresa foi marcado como tendo um sentimento global positivo (31%-21%), sendo também o maior *cluster* do modelo (37% dos *tweets*).

Música Burger King Trap House – O segmento com a música foi marcado como tendo um sentimento global negativo (22%-28%)

1.11.2. Modelos McDonald's

MODELO 1					
Cluster	*Sentimento*	Frequência	Total de *tweets* do *cluster*	Percentagem	*Sentimento* do *Cluster*
Produtos 1 (Fries, breakfast, tea)	Negativo	5828	33318	17,5%	Positivo
	Neutro	17960		53,9%	
	Positivo	9530		28,6%	
Sensações	Negativo	5765	22312	25,8%	Negativo
	Neutro	11526		51,7%	
	Positivo	5021		22,5%	
Produtos 2 (Happy meal, coffee, cream)	Negativo	12358	62771	19,7%	Positivo
	Neutro	31180		49,7%	
	Positivo	19233		30,6%	
Produtos 3 (Chicken nuggets)	Negativo	18922	82725	22,9%	Positivo
	Neutro	42413		51,3%	
	Positivo	21390		25,9%	
Produtos 4 (lemonade, strawberry), Serviço (Order)	Negativo	12060	60301	20,0%	Muito Positivo
	Neutro	26043		43,2%	
	Positivo	22198		36,8%	
Produtos 5 (French fires)	Negativo	12166	54754	22,2%	Neutro
	Neutro	30368		55,5%	
	Positivo	12220		22,3%	

Tabela 20 – *Sentiment Analysis* para os *clusters* do modelo 1 para o McDonald's

Produtos 1 (Fries, breakfast, tea) – Neste *cluster* os produtos/serviços batatas fritas, chá e pequeno-almoço atribuem um sentimento positivo ao *cluster* (29%-17%)

Sensações – O *cluster* que identifica sentimentos/sensações apresenta um *sentimento* negativo (23%-26%)

Produtos 2 (Happy meal, coffee, cream) – Neste *cluster* os produtos happy meal, café e natas atribuem um *sentimento* positivo ao *cluster* (31%-19%)

Produtos 3 (Chicken nuggets) – O produto frango atribui um sentimento geral positivo ao *cluster* (26%-23%)

Produtos 4 (lemonade, straberry), Serviço (Order) – os produtos de limonada e constituídos por morango, bem como o serviço de atendimento aos clientes atribuem um sentimento muito positivo a este segmento (37%-20%)

Produtos 5 (French fires) – No McDonald's o segmento que identifica o produto batatas fritas tem um *sentimento* neutro (arredondando às unidades 22,3%-22,2%)

MODELO 2					
Cluster	***Sentimento***	**Frequência**	**Total de *tweets* do *cluster***	**Percentagem**	***Sentimento* do *Cluster***
Job/minions	Negativo	11337	55585	20,4%	Positivo
	Neutro	29263		52,6%	
	Positivo	14985		27,0%	
?	Negativo	3764	18999	19,8%	Positivo
	Neutro	8970		47,2%	
	Positivo	6265		33,0%	
Produtos 1 (Big Mac, Happy Meal)	Negativo	3041	14495	21,0%	Positivo
	Neutro	7824		54,0%	
	Positivo	3630		25,0%	
Produtos 2 (Happy Meal), Experiência (primeira vez no McDonald's)	Negativo	4501	22472	20,0%	Positivo
	Neutro	10580		47,1%	
	Positivo	7391		32,9%	
Produtos 3 (coffee)	Negativo	7674	23406	32,8%	Muito Negativo
	Neutro	12434		53,1%	
	Positivo	3298		14,1%	
?	Negativo	610	14970	4,1%	Muito Positivo
	Neutro	2616		17,5%	
	Positivo	11744		78,5%	
Produtos 4 (chicken, cream)	Negativo	22829	98140	23,3%	Positivo
	Neutro	50058		51,0%	
	Positivo	25253		25,7%	
KFC	Negativo	8517	44285	19,2%	Positivo
	Neutro	24257		54,8%	
	Positivo	11511		26,0%	
Produtos 5 (French fries, breakfast)	Negativo	4826	23829	20,3%	Positivo
	Neutro	13488		56,6%	
	Positivo	5515		23,1%	

Tabela 21 – *Sentiment Analysis* para os *clusters* do modelo 2 para o McDonald's

Job/minions – O *cluster* identifica o trabalhar no McDonald's e os artigos promocionais dos minions oferecidos nos produtos Happy Meal. Existe um *sentimento* positivo associado (27%-20%).

? - Neste *cluster* não é possível discernir um tema ou conjunto de temas.

Produtos 1 (Big Mac, Happy Meal) – Os produtos Big Mac e Happy Meal têm um sentimento positivo associado (33%-20%).

Produtos 2 (Happy Meal), Experiência (primeira vez no McDonald's) – O produto Happy Meal e a experiência de visitar o McDonald's pela primeira vez representam um segmento positivo (33%20%).

Produtos 3 (coffee) – O café, identificado neste segmento apresenta uma conotação muito negativa. O que poderá dar sinais de necessidade de alterações ao produto em si.

? - Acrescentando ao facto de não ser possível discernir um tema ou conjunto de temas deste *cluster*, é classificado como Muito positivo com uma proporção muito desigual de positivos e negativos (78%-4%)

Produtos 4 (chicken, cream) - Os produtos frango e natas têm um *sentimento* positivo associado (33%-20%)

KFC – o segmento que identifica o concorrente KFC tem um *sentimento* positivo associado (26%-19%)

Produtos 5 (French fries, breakfast) – O *cluster* que identifica os produtos batatas fritas e pequeno-almoço têm um *sentimento* positivo associado (23%-20%)

MODELO 3					
Cluster	*Sentimento*	Frequência	Total de *tweets* do *cluster*	Percentagem	*Sentimento* do *Cluster*
Produtos 1 (Chicken, Ice cream)	Negativo	16125	88620	18,2%	Positivo
	Neutro	48388		54,6%	
	Positivo	24107		27,2%	
Produtos 2 (Big Mac, coke)	Negativo	3130	30122	10,4%	Muito Positivo
	Neutro	12035		40,0%	
	Positivo	14957		49,7%	
Job/Interview	Negativo	2728	12727	21,4%	Positivo
	Neutro	6697		52,6%	
	Positivo	3302		25,9%	
French Fries	Negativo	7259	31599	23,0%	Negativo
	Neutro	17576		55,6%	
	Positivo	6764		21,4%	
Serviço (McDrive)	Negativo	3086	23556	13,1%	Muito Positivo
	Neutro	8736		37,1%	
	Positivo	11734		49,8%	
?	Negativo	28036	101840	27,5%	Negativo
	Neutro	52078		51,1%	
	Positivo	21726		21,3%	
Experiência (primeira vez no McDonald's)	Negativo	6735	27717	24,3%	Positivo
	Neutro	13980		50,4%	
	Positivo	7002		25,3%	

Tabela 22 – *Sentiment Analysis* para os *clusters* do modelo 3 para o McDonald's

Produtos 1 (Chicken, Ice cream) – O segmento com produtos de frango e gelados obtém um *sentimento* positvo (27%-18%)

Produtos 2 (Big Mac, coke) - O segmento com produtos de Big Mac e Coca-Cola obtém um *sentimento* muito positvo (50%-10%)

Job/Interview – O *cluster* que aborda o tema de entrevistas e trabalho no McDonald's tem um *sentimento* positivo associado (26%-21%)

French Fries – Neste modelo o segmento das batatas fritas obtém um *sentimento* negativo (21%-23%)

Serviço (McDrive) – O *cluster* referente ao serviço McDrive obtém um *sentimento* muito positivo (50%-13%)

? - Acrescentando ao facto de não ser possível discernir um tema ou conjunto de temas, é o maior *cluster* do modelo (32%) e é classificado como negativo (21%-28%)

Experiência (primeira vez no McDonald's) - a experiência de visitar o McDonald's pela primeira vez representam um segmento positivo (25%-24%)

1.11.3. Precisão do Modelos de *Sentiment Analysis*

Foi recolhida uma amostra estratificada proporcional por cada classe de *sentimento* para se medir a precisão do algoritmo BoW (Bernardo & Henriques, 2014), representada na tabela 23. Tal foi feito uma vez que a maioria dos *tweets* utilizados na modelação têm um *sentimento* neutro associado (aproximadamente 50% para as duas empresas). De modo a fazer uma classificação "manual" do *sentimento*, os *tweets* foram lidos e a interpretação humana classificou-os enquanto positivos, negativos ou neutros.

	n= 650	n= 1130
	Burger King	McDonald's
Negativo	23,5%	23,4%
Neutro	50,3%	48,8%
Positivo	26,2%	27,8%

Tabela 23 – Representatividade amostral

Foi recolhida uma amostra de 650 *tweets* para o Burger e 1130 *tweets* para o McDonald's. Posteriormente, comparou-se o sentimento atribuído pelo algoritmo BoW com o sentimento real dos *tweets*. As seguintes matrizes de confusão permitem compreender a distribuição de precisão dos modelos para as duas empresas:

		Amostra Burger King n=650			
		Sentimento Real			
		Negativo	Neutro	Positivo	Total
Sentimento Modelo	Negativo	68%	18,4%	13,6%	100%
	Neutro	10,5%	72,3%	17,17%	100%
	Positivo	10,6%	27,5%	61,97%	100%
	Frequência	153	327	170	650

Tabela 24 – Matriz de confusão Burger King

		Amostra McDonald's n=1130			
		Sentimento Real			
		Negativo	Neutro	Positivo	Total
Sentimento Modelo	Negativo	61,7%	36,4%	1,9%	100%
	Neutro	6,7%	87,1%	6,2%	100%
	Positivo	6,1%	39,1%	54,8%	100%
	Frequência	219	700	211	1130

Tabela 25 – Matriz de confusão McDonald's

Para as amostras recolhidas, não fazendo distinção entre as classes positivo, negativo, ou neutro, e calculando a precisão global como o número de acertos sobre o total da amostra, o modelo para o Burger King obteve uma precisão de 69,1% e o do McDonald's uma precisão de 72%. Podemos

concluir que o algoritmo *Bag-of-Words* obtém uma precisão razoável na classificação do sentimento dos *tweets* tendo em conta a dimensão reduzida do *text*o dos mesmos (Bernardo & Henriques, 2014).

1.12. ESCOLHA DOS MODELOS FINAIS E DISCUSSÃO

Tendo em conta os modelos obtidos, tanto de *Clustering* como de *Sentiment Analysis*, utilizando uma conjugação das duas análises pretendeu-se escolher um modelo para cada empresa que pudesse representar as temáticas abordadas pelos utilizadores do Twitter durante a janela temporal em que os dados foram obtidos, bem como o *sentimento* associado a cada um dos segmentos identificados.

1.12.1. Modelo Final Burger King

Assumindo os três modelos de *Clustering* obtidos para o Burger King, temos que o primeiro modelo obteve 4 *clusters* e em cada um deles foi possível identificar temas, sendo o RMS STD médio do modelo de 0,110. No caso do segundo modelo, que também obteve 4 *clusters*, em um dos *clusters* não foi possível identificar temas de forma tão clara apesar de ter *clusters* mais homogéneos (0,079). Por fim, o terceiro modelo obteve também 4 clusters sendo que dois deles identificam a música Burger King Trap House, e num deles não é clara a interpretação do tema do cluster. A homegeneidade média dos *clusters* é de 0,113.

Aplicando *Sentiment Analysis* aos três modelos de *Clustering* já referidos, obtém-se a polaridade de cada um dos segmentos. O algoritmo que atribuiu o sentimento aos *clusters* mostrou-se capaz de identificar sentimentos tanto positivos como negativos associados aos *clusters* dos dois modelos. Assim sendo, após a conjugação das duas análises o modelo escolhido para o Burger King foi o modelo 1, uma vez que apesar de o segundo modelo ter uma melhor homogeneidade, não apresenta todos os *clusters* com um bom grau de interpretabilidade, acontecendo o mesmo com o modelo 3, que para além da interpretabilidade também não é tão homogéneo como os dois primeiros modelos.

Modelo Burger King			
Nome	**Frequência**	**Percentagem**	***Sentimento***
Mcwhopper	3283	10%	Neutro
Música Burger King Trap House	18295	55%	Positivo
Produtos 1 (ice cream, shake)	7194	22%	Negativo
Produtos 2 (chicken)	4326	13%	Muito Positivo

Tabela 26 – Modelo final para o Burger King

O modelo do Burger king acaba por ser um misto de identificação de produtos da cadeia de restaurantes bem como de temas que causaram algum "burburinho" na rede social durante o período de extracção dos dados. O primeiro *cluster*, **Mcwhopper**, que representa 10% do total de *tweets* do modelo, identifica a proposta de criação de um hambúrguer conjunto que o Burger King fez ao McDonald's na tentativa de fazer "tréguas" com a empresa rival como forma de celebração do dia Internacional da Paz, comemorado a 21 de Setembro. O McDonald's acabaria por rejeitar a proposta publicamente em comunicado feito online por parte do presidente da empresa. O sentimento associado ao *cluster* é neutro, indicando que foi uma polémica que dividiu a audiência.

O segundo *cluster* identifica a música **Burger King Trap House**, interpretada por um músico norte-americano. É o maior *cluster* do modelo com 55% do total de *tweets*, e o sentimento associado ao segmento é positivo.

O terceiro *cluster* identifica produtos alimentares comercializados pela empresa. O segmento de **gelados e batidos** representam 22% do modelo e têm um *sentimen*to negativo associado. O *sentimen*to associado a estes produtos deverá levar a empresa a investigar os motivos que causam este negativismo. Recolhendo 3 *tweets* deste segmento para análise:

Tweet 1 - *"@BurgerKing I went to @McDonalds for a shake cause your machines are always broken*"
Tweet 2 – "*Why is Burger King s Icee machine always down?*"

A análise destes *tweets* leva-nos a apontar alguns constrangimentos nas estruturas que suportam o fornecimento dos produtos aos clientes em loja, tais como máquinas e processos. A empresa poderá identificar as unidades com este tipo de constrangimento com vista à correcção do problema.

O quarto *cluster* identifica novamente um produto comercializado pela empresa. Neste caso, e apesar de ser o menor *cluster* do modelo (13% do total de *tweets*) o segmento de **produtos de frango** obtém um *sentimen*to Muito positivo**.** Poderá ser uma oportunidade de a empresa alavancar outros produtos associando-os a estes por forma a reforçar ainda mais este segmento.

1.12.1. Modelo Final McDonald's

O mesmo processo foi aplicado para o McDonald's. Dos três modelos de *Clustering* obtidos o primeiro obteve um RMS STD de 0,107, o segundo modelo obteve 0,103 e o terceiro modelo obteve 0,119. O primeiro modelo foi o único a partir do qual foi possível obter uma interpretação definida de todos os *clusters* sendo que os restantes dois criaram *clusters* onde não foi possível identificar temas de forma eficaz.

Aplicando *Sentiment Analysis* aos modelos de *Clustering* obtidos, obtém-se a polaridade de cada um dos segmentos. O algoritmo que atribuiu o *sentimento* aos *clusters* mostrou-se capaz de identificar *sentimentos* associados aos *clusters* dos três modelos. Após a conjugação das duas análises o modelo escolhido para o McDonald's foi o modelo 1, uma vez que apesar de o segundo modelo ter uma melhor homogeneidade, não apresenta todos os *clusters* com um bom grau de interpretabilidade.

Modelo McDonald's			
Nome	**Frequência**	**Percentagem**	***Sentimento***
Produtos 1 (Fries, breakfast, tea)	33318	11%	Positivo
Sensações	22312	7%	Negativo
Produtos 2 (Happy meal, coffee, cream)	62771	20%	Positivo
Produtos 3 (Chicken nuggets)	82725	26%	Positivo
Produtos 4 (Lemonade, strawberry), Serviço (Order)	60301	19%	Muito Positivo
Produtos 5 (French fires)	54754	17%	Neutro

Tabela 27 – Modelo final para o McDonald's

Por sua vez o modelo do McDonald's identifica maioritariamente produtos. O primeiro *cluster* identifica o segmento de **batatas fritas, pequeno-almoço e chá**. Tem 11% do total de *tweets* e apresenta um sentimento positivo.

O segmento que identifica **Sensações** é o segmento mais pequeno, com 7% e tem uma polaridade negativa associada. Tal poderá estar ligado à má qualidade dos produtos da empresa, uma vez que são identificadas as palavras fat, feel e sick. Foram recolhidos alguns *tweets* deste segmento para análise:

Tweet 1 – "*I m feeling sick already. Early morning McDonalds wasn t a good idea*"
Tweet 2 – "*After I eat mcdonalds I feel sick*"

A análise dos *tweets* poderá querer dizer que produtos da marca causam mal-estar aos consumidores, sendo necessária uma análise cuidada relativamente ao tema identificado, possivelmente com vista à alteração dos menus dos restaurantes.

O segmento com os produtos **Happy Meal, café e natas** por outro lado tem uma polaridade positiva representando 20% do modelo.

O mesmo *sentimento* é encontrado no segmento **Nuggets de frango**, sendo o maior do modelo com 26% do total.

O *cluster* com os produtos **limonada, morangos e o serviço de atendimento nas lojas** com 19% do total de *tweets* do modelo, obtém uma polaridade muito positiva, podendo a empresa apostar na conjugação destes produtos com outros numa óptica de cross-selling, com o intuito de alavancar o crescimento de outros produtos.

Por fim, o último *cluster* identifica **batatas fritas,** desta vez sem outros produtos. O sentimento associado é neutro totalizando 17% dos *tweets*. Tal poderá indicar alguma divisão de opiniões entre os consumidores acerca das batatas fritas da empresa.

CONCLUSÕES

Após a análise dos resultados anteriormente apresentados, podemos afirmar que é possível encontrar padrões nos dados do Twitter com transferência real e interesse em termos de negócio para as empresas.

Foram criados modelos de *Clustering* hierárquico com o intuito de descobrir padrões interessantes num conjunto de *tweets* obtidos para o Burger King e McDonald's. Foram identificados segmentos (através de algoritmos de *Clustering* hierárquico) que indicaram a existência de temas abordados pelos utilizadores durante o período em que os dados foram extraídos. Sobre os segmentos obtidos foi aplicado um algoritmo de *sentiment analysis* (*Bag-of-Words*) com o intuito de compreender a polaridade associada aos temas encontrados.

De um modo geral, foi possível verificar que os modelos para as duas empresas conseguiram identificar produtos/serviços comercializados pelas empresas. Maioritariamente, o *sentimento* associado aos temas descobertos é positivo, identificando-se um segmento em cada modelo cuja polaridade é negativa. No caso do Burger King o segmento de gelados e batidos, onde as estruturas (máquinas e processos) que suportam a disponibilização dos produtos nas lojas não se encontram operacionais. Relativamente ao McDonald's o segmento negativo identifica sensações de mal-estar sendo referidas as palavras *fat, feel* e *sick* que indicam a indução de mal-estar aos consumidores destes produtos.

Concluindo, tendo em conta que os índices de satisfação dos clientes das duas empresas têm vindo a decrescer, as descobertas deste estudo sinalizam deficiências que as empresas poderão considerar em planos de acção para melhoria dos seus serviços e incremento da satisfação dos seus clientes.

LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Ao longo da elaboração do trabalho foram sendo encontradas limitações ao seu desenvolvimento. Uma grande limitação encontrada prende-se com a volumetria de dados, uma vez que para além da obtenção dos dados, foi necessário transformá-los bem como proceder ao seu armazenamento. A extracção de informação do Twitter foi morosa uma vez que estes foram sendo obtidos em tempo real. O verdadeiro desafio prendeu-se com o manuseamento da informação, isto é, no processamento dos *tweets* até ao ponto em que se encontraram prontos para serem analisados. Devido a um erro na extracção de dados para a empresa Burger King, detectado praticamente na fase final da janela definida para a recolha de dados, a proporção de dados obtidos foi bastante inferior face ao McDonald's. Tanto o pré-processamento dos dados como a modelação foram processos lentos que consumiram muitos recursos em termos de processamento.

São deixadas as seguintes recomendações para futuros estudos:

- Utilização do algoritmo *Bag-of-Words* em paralelo com um algoritmo de *machine learning*, como por exemplo um classificador de Bayes, com o intuito de avaliar os dois modelos e perceber qual obtém a melhor performance.

- Fazer uma análise da evolução do sentimento dos *tweets* ao longo do tempo, com o intuito de perceber como se comporta o sentimento dos utilizadores face à empresa ao longo dos vários períodos temporais (dia, semana, mês) e fazer *Clustering* com base nessas janelas temporais definidas.

- Utilização de outros algoritmos de *Clustering*, tal como o *K-Means,* com o intuito de perceber que diferenças de resultados os modelos apresentariam.

- Utilização de outro tipo de software para a modelação de *clusters*, como *Phyton* ou *R,* que permitam uma maior variedade na escolha dos algoritmos de *Clustering* e parametrização dos mesmos.

OBRAS CITADAS

Akshay, J., Xiaodan, S., Tim, F., & Java, A. (2007). Why We Twitter: Understanding Microblogging. *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (pp. 56-65). ACM.

Albright, R. (2004). Taming Text with SVD. *SAS Institute Inc.*

American Customer Satisfaction Index. (2015). *ACSI Restaurant Report 2015.*

Antonellis, Ioannis, & Gallopoulos, E. (2006). Explorint term-document matrices form matrix models in text mining. *arXiv preprint cs/0602076*.

Bates, A. (2015). Using Term Statistics to Aide in Clustering Twitter Posts. *Diss. University of Colorado at Colorado Springs*.

Bernardo, I., & Henriques, R. (2014). *A Era de um mercado social: A relação entre o Twitter e o Mercado Accionista.*

Berry, M., & Browne, M. (s.d.). *Understanding search engines: mathematical modelling and text retrieval.* Vol. 17 Siam.

Bhasin, K. (6 de Fevereiro de 2012). *13 Epic Twitter Fails By Big Brands*. Obtido de Business Insider: http://www.businessinsider.com/13-epic-twitter-fails-by-big-brands-2012-2?op=1

Bo, P., & Lee, L. (2008). Opinion mining and sentiment analysis. Em *Foundations and trends in information retrieval* (pp. 2.1-2, 1-135).

Carnegie Mellon University. (14 de Setembro de 2009). Distances between Clustering, Hierarchical Clustering. Pittsburgh, USA.

Choudhary, S., Singh, V., & Chakraborty, G. (2015). *Application of Text Mining on Tweets to Analyze Information about Type-2 Diabetes.* OK. US: Oklahoma State University.

Deerswester, S., Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1990). Indexing by latent semantic analysis. 391- 407.

Domingos, P. (s.d.). Mining Social networks for viral marketing. *IEEE Intelligent Systems 20.1*, pp. 80-82.

Dong, A. (2008). *The language of design: Theory and computation.* Springer Science & Business Media.

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: finding nuggets in mountains of textual data. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 398-401). ACM.

Facebook. (22 de Novembro de 2015). *Newsroom*. Obtido de Facebook: https://newsroom.fb.com/company-info/

Ghosh, S., Roy, S., & Bandyopadhyay, S. (1(4) de 2012). A tutorial review on Text Mining Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, p. 7.

Gomes, H., Neto, M., & Henriques, R. (2012). *Text Mining: Análise de sentimentos na classificação de notícias.*

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Clustering validity checking methods: part II. *ACM Sigmod Record*, 31(3), 19-27.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 464-472.

Hearst, M. (2003). *What is Text Mining.* UC Berkeley: SIMS.

Hotho, A., Nürnberger, A., & PaaB, G. (2005). A Brief Survey of Text Mining. *Ldv Forum Vol. 20 No.1.*

Huberman, B. A., Romero, D. M., & Wu, F. (2008). Social networks that matter: Twitter under the microscope. *SSRN 1313405*.

Ifrim, G., Shi, B., & Brigadir, I. (2014). Event Detection in Twitter using Aggressive Filtering and Hierarchical Tweet Clustering. *SNOW-DC @ WWW.*

Kaplan, A., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 59-68.

Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision Support Systems*, 48(2), 354-368.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language*, 5(1) 1-167.

Mahesh, T., Suresh, M., & Vinayababu, M. (2010). Text Mining: Advancements, challenges and future directions. *Internation Journal of reviews in Computing, 3*, 61-65.

Maimon, O., & Lior , R. (2005). *Data mining and knowledge discovery handbook.* New York: Springer.

Matsubara, Takashi, E., Martins, C. A., & Monard, M. C. (2003). *Pretext: Uma ferramenta paea pré-processamento de textos utilizando a abordagem bag-of-words.* São Carlos: Instituto de Ciencias Matematicas e de Computação.

Miner, G. (2012). Practical text mining and statistical analysis for non-structured text data applications. *Academic Press*.

Mosley Jr, & Roosevelt, C. (2012). Social media analytics: Data mining applied to insurance Twitter posts. *Casualty Actuarial Society E-Forum.* Winter 2012 Volume 2.

Mudinas, A., Zhang, D., & Levene, M. (2012). Combining lexicon and learning based approaches for concept-level sentiment analysis. *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining.* (p. 5) : ACM.

Mullen, T., & Collier, N. (July de 2004). Sentiment Analysis using Support Vector Machines with Diverse Information Sources. *EMNLP*, pp. Vol 4, 412-418.

Murugesan, A. K., & Zhang, B. J. (2011). A New Term Weighting Scheme For Document. *In 7th Int. Conf. Data Min.(DMIN 2011-WORLDCOMP 2011).* Las Vegas, Nevada, USA.

Nardi, B., Schiano, D., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM*, 47(12), 41-46.

Pagolu, Murali, Chakraborty, & Goutam. (2004). *Taming Text with the SVD.* Oklahoma State University.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up: sentiment classification using machine learning techniques. *Proceedings of the ALC-02 conference on Empirical methods in natural language processing.* Association for Computational Linguistics: Volume 10 (79-86).

Pedersen, T., & Bruce, R. (1997). *Unsupervised text Mining.* Department of Computer Science and Engineering, Southern Methodist University.

Roul, Kumar, R., Devanand, O. R., & Sahay, S. K. (2014). *Web document clustering and ranking using tf-Idf based Apriori Approach.* 1406 5617: arXiv preprint arXiv.

Sagiroglu, S., & Sinanc, D. (2013). Big Data: A review. In Collaboration Technologies and Systems (CTS). (pp. 42-47). International Conference on IEEE.

Sanders, A., & DeVault, C. (2004). *Using SAS® at SAS: The Mining of SAS Technical Support.* SUGI 29.

SAS Institute Inc. (2001). *SAS CUSTOMER SUPPORT.* Obtido de http://support.sas.com/: http://support.sas.com/documentation/onlinedoc/miner/em43/spsvd.pdf

Statista. (22 de Novembro de 2015). Obtido de Statista: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM*, (pp. 178-185).